Article:
Read, J and Dang, TNY orcid.org/0000-0002-3189-7776 (2022) Measuring depth of academic vocabulary knowledge. Language Teaching Research. ISSN 1362-1688

https://doi.org/10.1177/13621688221105913

Measuring Depth of Academic Vocabulary Knowledge

John Read
University of Auckland

and

Thi Ngoc Yen Dang
University of Leeds

Pre-publication version

Citation of published version:

Read, J. & Dang, T.Y.N (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research*. Online First publication, July 11, 2022.
https://doi.org/10.1177/13621688221105913

Measuring Depth of Academic Vocabulary Knowledge

Abstract

In L2 vocabulary studies there is continuing interest in tests of depth of vocabulary knowledge, measuring various aspects of word knowledge other than just the form—meaning that is the focus of breadth (or size) tests. This study aimed to explore new formats that could be used as diagnostic tools for assessing depth of academic vocabulary knowledge. The participants were 222 first-year students taking an EAP programme at a university in Vietnam. The depth measure was a newly developed test of a sample of words from Gardner and Davies' (2014) Academic Vocabulary List, including sub-tests on receptive knowledge of synonyms, collocations and word parts. The participants also took the Updated Vocabulary Levels Test as a breadth measure, as well as a translation test. Results showed that most of the students had a relatively limited knowledge of general English vocabulary. There was a moderate correlation between the breadth and depth tests. The students had the best knowledge of synonyms, followed by word parts and then collocations. The collocation and word parts sub-tests included a Not Sure option to discourage blind guessing, and analyses of responses to this option offered useful insights into the students' test-taking behaviour.

Key words:

vocabulary testing; academic word knowledge; vocabulary depth; Updated Vocabulary Levels Test; university students in Vietnam

I. Introduction

The development and validation of vocabulary tests has been a very active area of research publication in the past ten years or more, but a number of concerns have been expressed about the narrow basis on which most of this work has been conducted. In a new position paper, Schmitt, Nation and Kremmel (2020) build on earlier criticisms by scholars such as Read (2013) and Schmitt (2014) to present an agenda for the way forward to better quality second language (L2) vocabulary assessment.

The essence of the argument is as follows. The field of vocabulary assessment has been dominated until recently by a small number of tests, notably the Vocabulary Levels Test (VLT) (Nation, 1983) and the Vocabulary Size Test (VST) (Beglar, 2010; Nation, 2012). Being freely accessible on the web, these tests have been used for a variety of assessment purposes with learners of English in all kinds of educational contexts worldwide. However, the modern theory of test validity holds that validation relates to the interpretation of the scores when a test is administered for a defined assessment purpose to a specific population of learners (Chapelle, 2012). Schmitt et al. (2020) go on to advocate more critical scrutiny of existing vocabulary tests and a more professional approach to developing new ones, so that they meet current standards of test design and validation.

The present study is an exploratory investigation to demonstrate how some principles of this approach can be implemented in testing depth of academic vocabulary knowledge. It involves the trialling of a new test format which builds on the word associates principle and includes some innovative item types for a specific population of learners. The writing of the test material draws on Gardner and Davies's (2014) Academic Vocabulary List (AVL). Perhaps most significantly, the test has been designed for a particular population: university students in Vietnam who are preparing to study through the medium of English, even though their vocabulary knowledge is thought to be limited. English-medium instruction has become a widespread phenomenon in many other countries as well where English has traditionally been a foreign language, with low levels of achievement from years of compulsory study of the language in schools (Hyland & Shaw, 2016). Thus, this kind of test is potentially of broader interest in such educational contexts.

II. Literature review

1. Breadth of vocabulary knowledge

Underlying work on L2 vocabulary assessment since the 1980s has been a distinction, commonly attributed to Anderson and Freebody (1981), between breadth and depth of vocabulary knowledge. Breadth can be conceived as the number of words that a learner (or user) knows in some sense. First language (L1) reading researchers have long recognised the critical role of vocabulary knowledge in reading comprehension ability (for a review, see Pearson, Hiebert & Kamil, 2007) and this has led to considerable interest in estimating the vocabulary size of native speakers at various ages through childhood and into adult life (Nation & Coxhead, 2021).

The first widely used measure of vocabulary breadth for L2 learners was Nation's (1983) VLT, which was originally intended as a diagnostic tool for classroom teachers. It included words across four discontinuous frequency levels: 2000, 3000, 5000 and 10,000, as well as a sample of academic words. As the VLT came to be widely used for research and pedagogical purposes, Schmitt et al (2001) conducted a large-scale international validation study of a revised version of the test, with a larger sample of words at each of the five levels, in the interests of more reliable measurement.

A more substantial development was the VST (Beglar, 2010; Nation, 2012), which was conceived as a true size test which could be administered both to L2 learners and native speakers. The original test covered words up to the 14th 1000 frequency level, and presented the target words in a multiple-choice format. In the last ten years the VST has become a family of tests, with some versions covering a higher and lower word frequency range and bilingual versions which present word definitions in particular L1s (Elgort, 2013; Nguyen & Nation, 2011; Zhao & Ji, 2016).

Most recently, Webb, Sasao, and Ballance (2017) have developed an Updated Vocabulary Levels Test (UVLT). The UVLT covers the 1st 5000 most frequent word families of English in a continuous series of five levels and excludes academic vocabulary as a separate category. It covers the higher frequency words that foreign language learners are most likely to know and need.

As scholars such as Stoeckel et al. (2021) have pointed out, size and levels tests should be seen as two distinct ways to measure breadth of knowledge. Size tests seek to

estimate the total number of words that an individual knows, often for research purposes, whereas levels tests assess mastery of particular higher-frequency vocabulary levels for pedagogical purposes such as selecting materials and setting vocabulary learning goals. Research has consistently shown that knowledge of high-frequency words is essential for comprehension (Dang & Webb, 2020; Schmitt & Schmitt, 2014), but a large number of learners in Vietnam as in other EFL contexts have insufficient knowledge of these words (e.g., Dang, 2020; Nguyen & Webb, 2017; Sun & Dang, 2020). Thus, a levels test was more appropriate for the aims of the present study.

It should also be noted that the VLT, UVLT and VST all use selected-response test items (word-definition matching and multiple-choice), which assess recognition of meaning and are subject to the effects of guessing. There is increasing evidence that meaning-recall items, which require the test-takers to supply the meaning of the target words, are more reliable and correlate better with measures of reading comprehension (McLean et al., 2020; Stoeckel et al., 2021, Zhang & Zhang, 2020).

2. Depth of vocabulary knowledge

This brings us to the concept of depth of vocabulary knowledge. Measures of breadth need to include reasonably large samples of target words to provide a reliable basis for their size estimates, and thus the test items focus on the learners' ability simply to link the form and meaning of L2 words, using formats like word – definition matching, multiple-choice, translation, and self-report (Yes/No). Although such tests have excellent measurement properties and correlate remarkably well with measures of reading comprehension and overall language proficiency (Alderson, 2005; Milton, 2013), establishing the form-meaning link is just the first step in developing word knowledge.

The most influential framework for the components of vocabulary knowledge is that of Nation (2013, p. 49), which can be summarised as follows:

- **Form**

  How is the word spelled? and pronounced?

  What parts does it have?
- **Meaning**

  What is its central concept and range of meanings?

  What other words are associated with it?

- **Use**

  What patterns of grammar and collocation does it fit into?

  How is it used appropriately?

Nation's full table also distinguishes between receptive and productive knowledge, which in the context of vocabulary testing has most commonly meant in practice the distinction between *recognising* the correct form or meaning of a target word and being able to *recall* it.

In terms of designing measures to assess depth of knowledge, there have been three main approaches (Read, 2004). One has been to focus on how well the *meaning* of the target word is known, including the learner's level of confidence in their knowledge, using instruments such as the Vocabulary Knowledge Scale (VKS) (Wesche & Paribakht, 1996). The VKS targets primarily Nation's "central concept" of the target word.

The second approach addresses Nation's second meaning question above: "What other words are associated with it?" The concept of individual words being incorporated into a growing lexical network in the learner's mind has featured prominently in the work of Meara and his colleagues (Meara, 1997, 2009), who have explored ways of characterising the overall state of the mental lexicon through word association techniques. From this perspective, depth of vocabulary knowledge is conceived as a densely structured lexical network, rather than being a property of individual words. The concept of word association was also the inspiration for Read's (1993, 1998) word associates format, which focuses on the ability of test-takers to select words that have semantic or collocational relationships with specific target words. Read's pioneering work, numerous other researchers have investigated the use of the format for various research and pedagogical purposes in a number of adapted forms (for a review, see Zhang & Koda, 2017).

Nevertheless, in research studies comparing breadth and depth of knowledge, the two original tests (WATs) are still often taken to be the standard measures of vocabulary depth. Qian (1999) initiated a simple research design in which Read's (1998) WAT is administered, along with a test of vocabulary breadth (the VLT; Nation, 1983; Schmitt et al., 2001), and a reading comprehension test. Qian's trend-setting study and the numerous others which have adopted the same basic design (e.g., Akbarian, 2010; Zhang & Yang, 2016) have consistently shown that the two vocabulary measures are quite strongly correlated, and typically in a regression analysis, the VLT has been the main predictor of the reading scores, with the WAT making a significant, but small additional contribution.

The third approach to assessing vocabulary depth identified by Read (2004) is to go beyond meaning and word associations to test other components of Nation's word knowledge framework. In a comprehensive review article, Schmitt (2014) identified work on learners' knowledge of three of the components (among others): words with multiple meanings (e.g., Verhallen & Schoonen, 1993); derived forms of words and word parts (Schmitt & Zimmerman, 2002; see also Sasao & Webb, 2017); and knowledge of collocations (Eyckmans, 2009; Gyllstad, 2009; see also Nguyen & Webb, 2017). Nation himself recommends (2013) that classroom teachers should design tests that selectively target aspects of word knowledge that fit their current vocabulary teaching objectives.

However, there have been a number of studies which have included measures of multiple components of vocabulary knowledge. Schmitt (1998) tracked the longitudinal development of four knowledge components of 11 general academic words, using an intensive interview procedure with just three postgraduate students. Webb (2005, 2007) pioneered the use of multiple individual tests in studies on initial vocabulary learning using pseudowords to measure which aspects of target word knowledge were acquired from the experimental treatments: orthography, meaning and form, grammatical functions, syntax, and association. Each aspect was tested through both recognition and recall of the target lexical forms and meanings. These two researchers found that the acquisition of the various knowledge components occurred in parallel, with none being markedly easier to acquire than others. However, Chui's (2006) study which measured academic vocabulary knowledge of university students in Hong Kong showed that knowledge of collocations and derived forms lagged behind the ability to identify the word class and to recall word meaning.

More recently, González-Fernández and Schmitt (2019) undertook a cross-sectional study of Spanish-speaking learners of English with a range of proficiency levels to model the acquisition of four components of vocabulary knowledge: the form-meaning link, derived forms, multiple meanings, and collocates. Knowledge of 20 target words was measured both "receptively" (through recognition-type items) and "productively" (through recall-type items). The recall measures were always more difficult than the recognition items. The correlations among the component measures were consistently high, ranging from .70 to .95, and the component measures also correlated strongly (.76 to .90) with vocabulary breadth as measured by the VLT (Schmitt et al., 2001). In terms of the individual components, the form-meaning link was the best-known type of receptive knowledge. However, somewhat contrary

to expectations, recall of the form-meaning link, multiple meanings, and derived forms were all more difficult than recall of collocates. The authors suggest that this result reflected at least partly the design of the respective measures used for these components.

3. Comparing breadth and depth

In his comprehensive review of the research on breadth and depth of vocabulary knowledge, Schmitt (2014) found it difficult to draw any clear-cut conclusions about the relationship between the two. This reflected the lack of a theoretical basis for defining the concept of depth as something truly distinct from vocabulary size, as well as the multiplicity of often poorly validated tests to measure various components of vocabulary knowledge operationally. Schmitt makes a useful distinction between empirical investigations of the nature of vocabulary knowledge, such as those just reported, and pedagogical initiatives to assess learners' vocabulary knowledge in the school and classroom context. The present study falls more into the latter category.

One initiative that is somewhat similar to ours is the project by Ishii and Schmitt (2009) to develop a multi-part measure of English vocabulary knowledge for Japanese university learners, taking account of typical problems such learners encounter. The instrument consisted of four components: a test of vocabulary size, using a bilingual word – definition matching format; a test to identify two different meanings of a word; a test of derived forms of target words; and a test to distinguish appropriate sentence contexts for words which were near-synonyms. Apart from the first test, which covered a range up to the 6000-word level, the other three measures were based on a sample of target words from the most frequent 2000 lemmas in the British National Corpus. The authors established scoring norms for learners with different vocabulary sizes, to help teachers to identify those whose scores did not follow the expected pattern of development for the other three measures.

Thus, there is essentially a twofold purpose for the present study, first to explore the design of new measures of vocabulary depth and secondly to use the tests to diagnose the academic vocabulary knowledge of a specific population of university students preparing for an English-medium undergraduate programme. The focus of this study was on receptive knowledge of high-frequency academic vocabulary these students would encounter in their academic textbooks.

III. Method

1. Participants

The participants were 260 native speakers of Vietnamese, who had received their whole prior education in that country. They typically had begun studying English as a compulsory subject in lower secondary school when they were around 11 years old. They were recruited from a one-year compulsory English language course at a university in a major city in Vietnam. After this course, they were going to study 40% of their academic subjects through English medium instruction (EMI). Based on their scores on the university's entrance English exam and the Vietnamese Standardised Test of English Proficiency (VSTEP)[1],the students had been classified into mainstream and fast track streams. The mainstream students, who were estimated to be at A2 level on the Common European Framework of Reference (CEFR), were taking 10 periods of English per week, whereas the fast-track students were more at B1 level and studied for 20 periods per week. The students participated in the study during their regular English periods, which lasted 50 minutes each. There were 31 class groups altogether, comprising 28 groups in the mainstream and another 3 in the fast track. The total number of students in the groups was 260, but only 222 of them completed all of the tests and were included in the analyses. Informal consultation with content lecturers and past students revealed that reading academic texts was a key task in the participants' EMI programmes. As knowledge of academic vocabulary is essential for comprehension of academic texts (Coxhead, 2018), the test developed in the present study aimed to measure knowledge of academic vocabulary.

2. Instruments

a. Vocabulary Breadth Test

We used the UVLT (Webb et al, 2017) to measure breadth of vocabulary knowledge in this study. Like the original VLT, the UVLT uses a matching format (see Figure 1). At each frequency level, there are 30 test items (definitions to be matched with the correct target word), with a total score of 150. We used a paper version of the test, delivered as part of the participants' English classes to help ensure that the students took the test seriously.

	avoid	contain	murder	search	switch	trade
have something inside						
look for						
try not to do						

Figure 1: Sample item set from the Updated Vocabulary Levels Test

b. Depth Test

The Depth Test was developed specially for this study. Its purpose was to measure how well the students knew high-frequency academic vocabulary in English, beyond being able to match the words with a synonym or short definition. In particular, this test of depth aimed to measure knowledge of synonyms, collocations, and word parts of known words. These aspects were chosen because they each represent one component of Nation's (2013) vocabulary knowledge framework: form, meaning, and use.

Selecting test items

For this test, high-frequency academic vocabulary was defined as the 1st 200 lemmas in Gardner and Davies's (2014) Academic Vocabulary List (AVL). There are several reasons for choosing this list. First, it was compiled from a 120-million-word academic sub-corpus of the Corpus of Contemporary American English (COCA). This sub-corpus is primarily composed of academic journal articles from various disciplinary areas. Second, the AVL uses the lemma rather than the word family as the unit for classifying word forms. The lemma is a base form (*govern*) plus its inflections (*governs, governing, governed*), whereas the word family includes a base form, its inflections but also its closely related derivations (e.g., *governor, governors, government, governments, governmental, governance*). Lemmatization helps to distinguish words according to their part of speech, their various meanings, and their derived forms, which are often not transparently related to the stem form of the word (Gardner & Davies, 2014). Choosing the lemma-based list also matched the purpose of a sub-test of our Depth Test, which measured the participants' word part knowledge. Finally, the AVL did not exclude general high-frequency words, provided that the words met the criteria for selection in terms of frequency, range, and dispersion in the academic sub-corpus. This reflects the fact that high-frequency general words could have distinct meanings and usages in academic texts (Dang, 2018; Hyland & Tse, 2007). It also takes into account the fact that

not all learners who start learning English for Academic Purposes have mastered high-frequency general words (Akbarian, 2010; Dang, 2020). The full AVL consists of 3000 "core academic" words (lemmas), but our focus was on the sub-set of high-frequency words (the 1st 200 AVL lemmas) that our participants were most likely to know.

Apart from the focus on the 1st 200 AVL lemmas, a second decision was to restrict the words to be used in the test – both the target words and others –mainly to the first 2000 word families in the BNC/COCA word frequency list (Nation, 2020), together with selected word families from the 3000-word level. Previous testing of the students' vocabulary knowledge, plus the judgement of their language lecturers, indicated that these were the English words which the students were mostly likely to know the meaning of. The whole point of a depth test is to explore whether learners' knowledge of a "known" word extends beyond the ability to recognise one of its meanings.

The logic for the selection of target words for the test was different from that for a test of breadth of vocabulary knowledge, where it is important to use a random selection procedure from the word frequency list in order to estimate the total number of known words. In a depth test, by contrast, once a particular domain of known vocabulary has been delimited (such as high-frequency academic vocabulary in this case), it is more appropriate to apply judgement to the choice of words. A depth test sets out to explore the quality of the learners' knowledge of the words in that domain. In addition, individual words vary in the number of their relatively frequent synonyms, distinct meanings, collocates, inflections and derived forms, so this limits the words that can fit in test items designed to assess a particular component of word knowledge, as discussed further below. One added contrast is that a depth test can increase the sampling of the specified domain by employing words from a similar frequency level as synonyms, collocates, distractors and contexts for the use of the target words.

Determining test formats

The starting point for the test design was the concept behind the word associates format, which was primarily based on paradigmatic (synonymous) and syntagmatic (collocational) relations between words. Whereas in word associates tests these two types of association have generally been incorporated in a single test item, they are assessed in separate parts of this Depth Test, to avoid confusion between the two types and to give more flexibility in assessing collocational relationships.

Part A of the Depth Test focused on synonyms. It uses a simple selected-response format, which presents two synonyms[2] and two distractors for each target word, as in these examples:

system	organization	structure	difference	position
develop	improve	climb	advance	depend

In the first example, the correct options *organization* and *structure*, are high-frequency AVL lemmas, as is one of the distractors (*difference*), whereas the other distractor *position* is from the 1000-word family level of the BNC/COCA list. In the case of *develop*, where *improve* and *advance* are the correct options, *climb* is a word at the 1000-word family level of the BNC/COCA list, *improve* is in the 1st 200 lemmas of the AVL list, and the other two options are less frequent AVL lemmas, but still within the 1000- and 2000-word family levels of the BNC/COCA list. Thus, apart from representing or not representing the target word, the four options were chosen from within the 1st 200 AVL lemmas and/or the 1000-and 2000-word family levels of BNC/COCA.

Part A consisted of 30 items in total. Of the 120 words used for the options, 27 came from the 1st 200 AVL lemmas and another 18 were less frequent AVL words. Analysis with the BNC/COCA lists revealed these 120 words were also high-frequency general words: 67 words (e.g., *agree, think, same*) were from the 1000-word level, 49 (e.g., *discuss, improve, condition*) were from the 2000-word level and 4 (*contest, custom, international, structure*) were from the 3000-word level.

Part B focused on collocations. An original item format was devised, involving whole phrases and short sentences to allow various types of collocations to be assessed. In addition, the response format was changed from the selected responses in Part A to a Yes/No response for each item, as in the examples in Figure 2, which have the correct answers indicated. The Not Sure option was added, and placed in the right-hand column rather than the middle one, in response to feedback from students who participated in a trial of the draft test. They preferred not to have a forced choice in cases where they were not sure of the correct answer. However, it should be noted that including the Not Sure option introduced a new source of variance to the test scores – willingness to use this option – as Stoeckel *et al.* (2016) showed in their investigation of the "I Don't Know" option in the VST.

	Yes	No	Not Sure
a. a low level	✓		
b. to write an article	✓		
c. a closed difference		✓	
d. the process looks loud		✓	
e. the population grows each year	✓		

Figure 2: Sample items for Part B of the Depth Test

The correct items were taken, in edited form, from the concordance lines and common collocates for each target word in the AVL database: https://www.wordandphrase.info/academic/frequencyList.asp. On the other hand, the incorrect items required more subjective judgement on the part of the first author, as item writer, as to what represented a plausible collocation that was not found in the database, and then the items were reviewed by the second author, as well as being checked for non-occurrence in a Google search and in the COCA. Of the 30 items in Part B, 21 were correct and 9 were incorrect.

Part C of the test was designed to measure knowledge of different forms of the target words. In keeping with the word parts component of Nation's (2013) framework, both inflected and derived forms of the words were included. The first format to be trialled for Part C required the test-takers to supply the correct word form, as in the sample item in Figure 3:

produce

a. the total _____ of rice this year

b. the oil _____ countries in Asia

c. the selling of various agricultural _____

Figure 3: Draft word part items for Part C of the Depth Test

The feedback from the trial participants was that this kind of constructed-response item was too challenging for them[3]. Therefore, the format was changed to one that was similar to that for Part B (Figure 4). As we pointed out in the literature review, this change in

format potentially meant the Part C results were less reliable and less indicative of the learners' ability to employ this knowledge in reading.

	Yes	No	Not Sure
produce a. the total producement of rice this year b. the oil producing countries in Asia c. the selling of various agricultural products			

Figure 4: Sample items for Part C of the Depth Test

There were 20 target words in Part C, yielding a total of 60 items. For individual words, there were one, two or three correct word forms.

c. Translation Test

A simple Translation Test was devised to provide an independent measure of the participants' knowledge of the 30 target words in Part A of the Depth Test. Participants were instructed to write the meaning of each word in their L1 (Vietnamese). The primary purpose of the Translation Test was to provide corroborating evidence that the test-takers had some knowledge of the meanings of the target words in Part A, rather than simply making uninformed guesses. As noted by Stoeckel *et al.,* (2021), the Part A format (like that of the UVLT) was more sensitive to partial knowledge of the target words, whereas the meaning-recall task in the Translation Test was a better indicator of word knowledge available for reading.

3. Procedure

The tests were administered to the participants by the second author during their English classes on two separate days in the same week. On Day 1, the participants completed the Depth Test (30-40 minutes). Then, they had a 50-minute break and completed the Translation Test (15-20 minutes). On Day 2, they completed the UVLT (30-50 minutes).

The tests could be objectively scored, with the exception of the Translation Test. In this case, the second author collated all the responses and classified them into lists of correct

or incorrect translations of the target words. The lists were reviewed by another rater, who was an experienced interpreter and translator. There was a high level of agreement between the two raters' judgements, as confirmed by Cohen's kappa ($\kappa = .91$, $p < .0005$). Drawing on feedback from the second rater, a final marking key was agreed for the test.

4. Analysis

The statistical analyses were performed using SPSS for Microsoft Windows (Release 23.0). First, reliability estimates were calculated for the three tests, along with descriptive statistics for the whole tests and for the parts of the UVLT and the Depth Test. Since the three parts of the Depth Test had different numbers of items, the scores were normalised to facilitate the comparison of the parts. Based on the UVLT scores, the test-takers were divided into three groups according to their mastery of the 1000- and 2000-word frequency levels, in order to be able to make some analyses according to their level of vocabulary knowledge.

To explore the relationship between breadth and depth of knowledge of the target words, correlations were calculated between the UVLT, the Depth Test and their parts. This was followed by several analyses to probe the validity of the three parts of the Depth Test. The items in Part A were compared with the corresponding items in the Translation Test as an independent measure of knowledge of the target words. In Parts B and C the test-takers' performance on the correct and incorrect items was compared. There was also a careful analysis of how the Not Sure option operated as a response to the different types of item in Parts B and C because, as Stoeckel *et al.* (2016) showed in their investigation of the comparable I Don't Know option in the VST, it introduced a new source of variance to the test scores: willingness to use this option.

IV. Results

1. Test Reliabilities

The reliability of the tests was estimated using Cronbach's alpha (Table 1). As is typically found with vocabulary tests composed of many items, the tests were highly reliable, with the exception of Part B of the Depth Test, which had a less satisfactory alpha of .75. This part is discussed further below.

Table 1: Reliability estimates of the tests

Test	alpha
Updated Vocabulary Levels Test:	.96
Depth Test	
Total Score:	.95
Part A:	.94
Part B:	.75
Part C:	.90
Translation Test	.94

2. Vocabulary breadth

A measure of the size of the participants' higher-frequency English vocabulary was obtained by means of the UVLT. Table 2 presents the descriptive statistics for the whole test and for each of the five subtests, representing the 1000- to 5000- word levels of the BNC/COCA word frequency lists (Nation, 2020).

Table 2. Descriptive statistics of the Updated Vocabulary Levels Test (*N*=222)

Level	Mean	SD
1000-word	26.66	3.89
2000-word	18.51	8.19
3000-word	11.22	8.31
4000-word	8.47	7.60
5000-word	6.84	7.42
Total score (k=150)	71.70	31.02

The overall mean score of 71.7 out of 150 words (47.8%) shows that the test was moderately difficult for these students, and to the extent that the target words constitute a representative sample of the 5000 most frequent word families, the mean gives an estimated average vocabulary knowledge of 2370 out of the most frequent 5000 words. As indicated by the standard deviation, there was a wide dispersion of scores, which is explored further below.

For the five individual frequency levels, we see the expected pattern of declining mean scores from the most frequent (1000) to the least frequent (5000) levels, reflecting the well-established finding that the likelihood of a word being known is strongly related to its frequency in the language (Read, 1988; Milton, 2009). Results of a one-way repeated measures ANOVA showed that the differences in the mean scores across the five levels were statistically significant (Wilks' Lambda = .02, F (5,217) = 2554.98, p <.001, η^2=.98). Post-hoc tests indicated that all of the differences in the individual mean scores were also significant.

Although the overall level of English vocabulary knowledge among the participants was relatively low, it is useful to divide them into sub-groups according to their mastery of the first three 1000-word frequency levels in the UVLT. The authors of the test (Webb et al. 2017) recommend that the cut score for mastery at these frequency levels should be set at 29/30 (97%) because this high frequency vocabulary accounts for such a high percentage of the running words in any text. However, this cut score is very stringent compared to those applied to the VLT by earlier researchers, such as Read (1988), 16/18 (88.9%); Schmitt et al. (2001), 26/30 (86.7%); and Xing and Fulcher (2007), 24/30 (80%). Arguably, a score of 29/30 makes too little allowance for measurement error in test-taker performance and the sampling of words from the frequency list.

From the perspective of criterion-referenced language testing, Brown and Hudson (2002, pp 265-268) argue that cut scores need to be set in relation to the purpose of the test. They consider that a cut point of 90% to be desirable for placement or diagnostic purposes, whereas 60% is more acceptable as a minimum level of achievement at the end of a course.[4] Since our purpose here is more diagnostic in nature, we have adopted a cut score of 27/30 (90%) to divide our participants into three sub-groups:

a) 82 learners who had not mastered even the1000-word level (<1K learners);
b) 93 learners who had mastered only the 1000-word level (1K learners)
c) 47 learners who had mastered the 1000-word level and at least the 2000-word level (2K+ learners).

Within the 2K+ group, 12 learners had mastered the 3000-word level, 7 the 4000-word level, and 6 the 5000-word level.

Table 3. Descriptive statistics of the Updated Vocabulary Levels Test for the three sub-groups

	<1K learners (*n*=82)		1K learners (*n*=93)		2K+ learners (*n*=47)	
UVLT level	Mean	SD	Mean	SD	Mean	SD
1000-word	22.66	3.72	28.59	1.04	29.81	0.40
2000-word	11.56	4.49	19.61	6.26	28.45	1.21
3000-word	5.52	4.77	11.01	6.74	21.55	5.82
4000-word	3.87	3.75	8.22	5.75	17.02	8.59
5000-word	2.41	3.23	6.61	6.07	15.02	8.31
Total score	46.02	16.05	74.04	21.00	111.85	21.29

The descriptive statistics were then calculated separately for the three groups, as set out in Table 3. The total scores show that the three sub-groups were very clearly differentiated in terms of vocabulary breadth. The figures highlight the point that the <1K group had not only failed to master the high-frequency words at the 1K level but knew just less than one-third of the 150 target words overall. By contrast, the small group of learners in the 2K group demonstrated on average a knowledge of 75% of the words in the test.

Thus, the results from the UVLT show that most of the participants had a reasonably small receptive knowledge of high-frequency English vocabulary, centred on the ability to recognise at least one meaning of most of the first and second thousand (1K & 2K) word families. The findings of the present study are consistent with previous research which shows that a considerable number of learners in Vietnam as well as in other EFL contexts have insufficient knowledge of high-frequency words (e.g., Nguyen & Webb, 2017; McLean & Stoeckel, 2021).

3. Depth of knowledge

We move now to the results of the Depth Test. Table 4 presents the descriptive statistics for the three parts and the test as a whole. The overall mean score was 78.9 out of 150 items (52.6%). Although the Depth Test happened to have the same number of items as the UVLT, it is not meaningful to compare the mean scores of the two tests because of the rather different designs and samples of words used in each one.

Table 4. Descriptive statistics of the Depth Test

	Raw scores		Normalized scores	
	Mean	SD	Mean	SD
Part A – Synonyms (*k*=60)	34.19	12.92	5.70	2.15
Part B – Collocations (*k*=30)	12.76	4.92	4.25	1.64
Part C – Word parts(*k*=60)	31.99	11.06	5.33	1.84
Total score – A+B+C (*k*=150)	78.93	25.10	--	--

It is more useful to compare the three parts of the Depth Test. Since Part B had a smaller number of items than Parts A and C, the mean scores are given in both raw and normalized form in Table 4. The normalized scores were calculated by dividing each raw score by the number of items in that part of the test and multiplying by 10. Thus, for example, a raw score of 42 in Part A yielded a normalized score of 7.0 (42÷60 x10).

Both raw and normalized mean scores show that Part A (Synonyms) was the least difficult part of the test, followed by Part C (Word families) and then Part B (Collocations). Results of a one-way repeated measures ANOVA showed that there were statistically significant differences in the mean scores across the three parts (Wilks' Lambda = .56, F (2,220) = 85.69, p <.001, η^2=.44). Post-hoc tests indicated that all the individual differences were significant.

The pattern of difficulty of the three test parts varied a little among the three sub-groups of participants defined above, as shown in Table 5. The <1K learners in particular performed differently from the other two sub-groups. For them, the Part B mean score was significantly lower than the other two, but their scores on Parts A and C were virtually the same. This may reflect their generally limited knowledge of the vocabulary. On the other hand, the two sub-groups with larger vocabulary sizes (1K and 2K+) showed the overall pattern, whereby knowledge of synonyms was significantly greater than knowledge of word family members, followed by knowledge of collocations.

Table 5. Descriptive statistics of the normalized Depth Test scores of each group of learners

Test Part	<1K learners (*n*=82)		1K learners (*n*=93)		2K+ learners (*n*=47)	
	Mean	SD	Mean	SD	Mean	SD
Part A – Synonyms	4.38	1.77	5.82	1.95	7.76	1.30
Part B – Collocations	3.60	1.65	4.17	1.43	5.55	1.26
Part C – Word parts	4.60	1.82	5.28	1.74	6.71	1.22

4. Relation between vocabulary breadth and depth

To explore the relationship between the learners' vocabulary breadth and their depth of academic vocabulary knowledge, Pearson correlations were calculated between the scores from the UVLT and the Depth Test, as presented in Table 6. The correlation between the total scores of the tests was .64, which indicated a moderate relationship accounting for 41% of the shared variance. It should be remembered here that the target words for the two tests were derived from different word frequency lists, although there was considerable overlap between the two lists. The correlations of the overall Depth Test scores with the individual frequency levels of the UVLT were comparable, apart from the more modest coefficient with the 1000-word level, which is likely to reflect the limited variance in the scores on the 1000-word level.

Table 6. Correlations between the UVLT scores and the Depth Test scores (*N* =222)

UVLT scores	Depth Test scores			
	Overall	Part A	Part B	Part C
1000-word level	.44	.48	.27	.32
2000-word level	.57	.57	.38	.45
3000-word level	.62	.60	.42	.52
4000-word level	.57	.56	.39	.46
5000-word level	.57	.56	.38	.46
Overall score	.64	.64	.43	.52

All correlations significant at the 0.01 level (2-tailed).

Looking at the three parts of the Depth Test, we see that the correlations of Part A with the UVLT are very similar to those for the Depth Test overall. We might expect Part A to correlate quite well with the UVLT because both are measuring knowledge of word meaning. On the other hand, the correlations with Part C are lower, and those with Part B are more modest again. Thus, the relative size of the correlations reflects the order of difficulty of the three parts of the Depth Test, as presented in Table 4, and also offers some evidence that the dimensions of word knowledge measured by Parts B and C may be distinct from what is assessed by Part A.

It is also interesting to note that, of the five frequency levels in the UVLT, it is the 3000-word level which produced the highest correlations with the Depth Test (both the parts and the overall score)[5].

5. *Depth Test, Part A (Synonyms) and the Translation Test*

The Translation Test included the 30 target words from Part A of the Depth Test. As previously stated, its primary purpose was to control for the possibility that there was a large element of guessing in their responses to the Part A items. For this purpose, Item Facility (IF) values (the proportion of test-takers answering an item correctly) were calculated for the two sets of items. The IFs for the Translation items covered a wide range, from .92 for *technology* to .17 for *indicate*, with a mean of .65 and a standard deviation of .22. In the case of the Part A items, IFs were calculated separately for the two correct associates in each item. Here again the range was wide, from .93 for *make* as an associate of *produce* to .17 for *regularity* as an associate of *pattern.*

A Pearson correlation of .77 was obtained between the IFs for the 30 Translation items and for the easier of the 30 associates of each Part A item (on the assumption that this associate expressed the more familiar sense of the target word, in cases where the two associates represented different senses). When the Part A IFs were re-calculated as the sum of the IFs for the two associates of each item, the correlation with the Translation IFs was just a little lower, at .75.

Thus, in most cases the difficulty level of the Translation items was comparable to that of the easier of the two associates in Part A, with a discrepancy in the IFs of no more than .15. Those items with a larger gap are shown in Table 7.

Table 7. Items in the Translation Test and Part A of the Depth Test with substantially different Item Facility (IF) values

Target Word	Translation IF	IFs of associates in Part A	
report	.87	.59 (describe)	.50 (announce)
control	.86	.61 (influence)	.53 (power)
tradition	.82	.55 (habit)	.38 (custom)
value	.80	.55 (benefit)	.42 (worth)
economic	.77	.48 (commercial)	.47 (financial)
produce	.68	.93 (make)	.86 (create)
indicate	.17	.42 (show)	.42 (suggest)

For the first five target words, most of the students gave a correct translation but were less successful in selecting the two associates in the Depth Test. This was probably a combination of having less knowledge of some of the associate words, and not being able to make a semantic connection between the target word and the associate, perhaps because the associate represented an unfamiliar sense of the word. On the other hand, for the last two target words the pattern was reversed: the test-takers were rather more successful at selecting the associates than providing an acceptable translation of the target word. The familiarity of the words used as associates for these two items may have played a role here.

6. Depth Test, Part B (Collocation)

The test-takers obtained the lowest mean score in Part B of the Depth Test, where they were asked to make a Yes/No judgement as whether each of a set of 20 expressions included a common collocation in academic English. They also had the option of a Not Sure response. Table 8 shows the response patterns for the five easiest and the five most difficult items.

Table 8: Responses to the Easiest and Most Difficult Items in the Depth Test, Part B

Items	Item Facility	Frequency of Responses		
		Yes	No	Not Sure
Easiest Items				
similar effects	.75	167	27	29
the global market	.68	150	33	38
a research project	.65	143	38	41
to establish a relationship	.64	142	34	46
the nature of human life	.64	142	48	31
Most Difficult Items				
the reaction of tradition	.23	110	51	61
the social century	.23	109	51	62
the research contains progress	.23	66	51	105
a straight strategy	.19	95	43	84
a basic council	.17	133	37	52

It is interesting to note that the first set of easier items were all acceptable collocations, with around two-thirds or more of the test-takers responding with Yes. There were relatively small and even numbers of No and Not Sure responses. On the other hand, the most difficult items all involved non-collocations, but less than a quarter of the test-takers rejected them with a No response. Not Sure responses were higher – and dramatically so in the case of *the research contains progress*. For three of the difficult items, half or more of the test-takers gave a Yes response.

We can also look at the overall response patterns of the test-takers to see the extent to which individuals made use of the Not Sure option. Table 9 shows that for the whole sample of participants this option was chosen 8.71 times for the 20 items, indicating a considerable degree of uncertainty as to whether each expression was acceptable or not. As might be expected, the frequency of the Not Sure option had a strong negative correlation with the

score on Part B and to a lesser degree on the overall score on the Depth Test. The table also shows these relationships separately for the three sub-groups according to their vocabulary knowledge. Students with a larger vocabulary size, especially the 2K+ learners, made much less use of the Not Sure option and the correlation with the Part B score was somewhat lower. However, there is no such trend in the correlations with the total Depth Test scores.

Table 9: Means and correlations of Not Sure responses to Part B of the Depth Test

	N	Not Sure Responses		Correlations	
		Mean	S.D.	Part B Total Score	Depth Test Total Score
Whole sample	222	8.71	7.44	-.832	-.638
<1K group	82	11.18	8.05	-.862	-.597
1K group	93	8.37	6.89	-.814	-.636
2K+ group	47	5.06	4.71	-.628	-.521

The pattern of individual usage of the Not Sure response in Part B is summarised in Table 10. The tallies covered the full range from 0 to 30 responses. Two-thirds of the test-takers selected Not Sure 10 times or fewer, with 25 of them not giving this response at all. On the other hand, a diminishing number were uncertain about the acceptability of most if not all of the 30 collocational expressions.

Table 10: Tallies of Not Sure responses by individual test-takers in Part B of the Depth Test

NS responses	**Number/percentage of test-takers**
0	25 (11.3%)
1-10	124 (55.9%)
11-20	50 (22.5%)
21-30	23 (10.4%) incl. 3 with 30 NS responses

7. Depth Test, Part C (Word parts)

The easiest items in Part C (see Table 11) demonstrated that the learners were familiar with inflected and derived forms of nouns in particular. In most cases the relative easiness of the target forms was comparable to the level of difficulty of the corresponding target word in the Translation Test. Three exceptions were *achievement*, *association* and *identification*, where the test-takers found it much easier to recognise the derived noun form than to translate the corresponding stem form, which was a verb. The discrepancy in difficulty between *performed* and *perform* is more difficult to explain.

In the case of the difficult items in Part C (Table 12), as was true in Part B they were all incorrect or inappropriate forms in the context of the sentence in which they occurred. For six of the items, the item facility was much lower than that of the corresponding target word in the Translation Test, indicating that most of the students were able to supply the meaning of the target word but not able to reject an erroneous derived form. On the other hand, only a minority of the test-takers could translate *determine, recognize, perform* and especially *specific* correctly, so that presumably helps to explain why many were not able to identify the incorrect forms in Part C.

Table 11: Easiest items in Part C of the Depth Test, compared to Translation Test results

Items	**Item Facility**	**Frequency of Responses**			**Translation Item Facility**
		Yes	**No**	**Not Sure**	**(target word)**
They admired his level of achievement.	.83	185	8	29	.55 (achieve)
There is a relationship between drugs and crime.	.80	179	22	21	.84 (relationship)
The doctor performed several tests.	.80	178	26	18	.42 (perform)
He acted as the reporter for the project.	.75	166	34	22	.87 (report)
They are researchers at the university.	.75	167	35	20	.78 (research)
Each culture has its own traditions.	.74	165	26	31	.82 (tradition)
The reports were sent to our manager.	.74	165	42	15	.87 (report)
The doctors formed their own association.	.73	163	14	45	.22 (associate)
They stay healthy in natural ways.	.73	163	20	39	.90 (nature)
His work involves the identification of new plants.	.72	160	23	39	.48 (identify)

Table 12: Most difficult items in Part C of the Depth Test, compared to Translation Test results

Items	Item Facility	Frequency of Responses			Translation Item Facility
		Yes	No	Not Sure	(target word)
They investigated the issue researchly.	.42	78	93	78	.78 (research)
She paid the amount determinated by the court.	.41	71	90	61	.42 (determine)
They studied a relatable problem.	.40	71	89	62	.84 (relationship)
He is a skilled performist.	.38	72	85	65	.42 (perform)
She received recognizement for her efforts.	.37	78	82	62	.46 (recognize)
The painting was valuated at $50,000.	.36	109	80	33	.80 (value)
We were able to specificate the equipment we wanted.	.31	81	68	73	.29 (specific)
The people in this town have strong cultural valuations.	.31	97	68	57	.80 (value)
My brother is an economician.	.26	117	58	47	.77 (economic)
There is a basical mistake in the calculation.	.26	126	57	39	.81 (basic)

The tallying of the Not Sure responses to the items in Part C (Tables 13 and 14) revealed similar patterns to those in Part B. The mean number of responses at 12.78 represents 21.3% of the total responses, which is somewhat lower than the 29% of Not Sure responses to Part B. According to Table 14, more than three-quarters of the test-takers used this response for fewer than a third of the test items.

The declining means across the three vocabulary size sub-groups and the strong negative correlations with the Part C total score provide clear evidence again that more vocabulary knowledge gave learners confidence in their judgements about whether the target inflected and derived forms were correct or not.

Table 13: Means and correlations of Not Sure responses to Part C of the Depth Test

	N	**Not Sure Responses**		**Correlations**	
		Mean	**S.D.**	**Part C Total Score**	**Depth Test Total Score**
Whole sample	222	12.78	13.05	-.811	-.703
<1K group	82	16.26	14.78	-.882	-.770
1K group	93	12.23	12.99	-.760	-.689
2K+ group	47	7.81	6.99	-.706	-.553

Table 14: Tallies of Not Sure responses by individual test-takers in Part C of the Depth Test

NS responses	**Number/percentage of test-takers**
0	33 (14.9%)
1-20	137 (61.7%)
21-40	39 (17.6%)
41-60	13 (5.9%) incl. 1 with 60 NS responses

V. Discussion

This study has had a twofold purpose: to explore some innovative item formats to assess the depth of vocabulary knowledge, while measuring the academic vocabulary knowledge of a particular student population.

1. Vocabulary knowledge of the examined student population

The participants' breadth of vocabulary knowledge was measured by means of the UVLT. The test samples from the 5000 most frequent word families in English in Nation's (2020) BNC/COCA lists, covering the high-frequency vocabulary in the language as well as some of the mid-frequency range. According to vocabulary researchers such as Nation (2006) and Schmitt and Schmitt (2014), mastery of high-frequency words up to around the 3000-word family level is the first priority for L2 learners, in order to deal with general communication and have a solid foundation for further lexical development. As previously mentioned, after the English language course the participants were going to study a number of their academic subjects through English-medium instruction, in which they would need to read academic texts in English. In our study, with 27 out of 30 as the criterion for mastery, more than a third of the participants had an incomplete knowledge of even the first 1000 words. On the other hand, only about 20 percent of them had mastered the 2000-word level or higher. The estimate of a mean knowledge of 2370 out of the 5000 words represented in the test is another indication that overall the students had a barely adequate knowledge of high frequency vocabulary.

It should be noted that the BNC/COCA lists comprise general rather than academic vocabulary and, unlike the original VLT, the updated version that we used does not include a sample of academic words. This is one reason that we adopted Gardner and Davies' AVL as the basis for our Depth Test, to focus more specifically on words that occur with high frequency in written academic texts. Gardner and Davies (2014, pp. 308-310) showed that many of the most frequent word families occur, often with distinct meanings, in academic discourse. Additionally, studies with EAP learners in different contexts have reported that not all learners had mastered high-frequency words when they started their EAP studies (e.g., Akbarian, 2010; Dang, 2020). At the time of our data collection, there was no standard vocabulary breadth test based on the AVL[6] and so the UVLT was the best option available to assess knowledge of higher frequency vocabulary in English. We would argue that the lack of mastery of this vocabulary revealed by our UVLT results provides evidence that most of the

students are likely to fall well short of achieving adequate coverage of the lexical content of the reading materials in their future English-medium courses, with consequent negative effects on their ability to comprehend the materials.

2. Evaluating the format of the Depth Test

The Depth Test is an experimental measure, designed as much to explore new test formats as to draw definite conclusions about the vocabulary depth of this population of students. The correlation with the UVLT scores showed a moderate relationship of .64 between breadth and depth, as measured in this study. This is within the range of correlations that have been obtained in previous studies of the relationship between these two dimensions (Schmitt, 2014), although comparisons are a little problematic because of the variety of measures used by different researchers.

As could be expected, Part A of the Depth Test (Synonyms) correlated best with the UVLT, since both assessed knowledge of word meaning. The test-takers were not offered the Not Sure option in this part of the test. In introducing the test, the administrator encouraged the test-takers to make thoughtful guesses, but to leave one or both options unmarked if they had no idea about which options were correct. A review of the responses indicates that they likely did not mark a response if they did not know which option(s) to choose, rather than guessing blindly. This is consistent with their liberal use of Not Sure in Parts B and C. It is also corroborated by the relatively strong correlations between the responses to Part A and the results of the Translation Test, which provided an independent, constructed-response measure of the learners' knowledge of the target words. This allays concern about the guessing factor in a format which involved selecting two of four possible responses.

The cases of discrepancy between the difficulty levels of words in Part A and the Translation Test were relatively few and apparently reflected either the students' lack of familiarity with the associates or their inability to make a semantic link between the target word and each associate.

Turning to Part B of the Depth Test, it is understandable that this was the most difficult part because it required knowledge not only of the individual words but their collocational possibilities. Knowledge of collocations is recognised as a source of difficulty, even for relatively advanced learners (Laufer & Waldman, 2011; Nguyen & Webb, 2017**).** The results showed that a majority of students were able to recognise common collocations

(e.g., *similar effects*, *the global market* and *a research project*), but they decisively rejected only one of the non-collocational expressions, *to produce culture*. They were less confident about judging the other incorrect items (see the examples in Table 8 above), which generally produced many more Not Sure responses. There was also uncertainty about some of the correct items, such as *the findings were very positive*, *a fixed exchange rate, the current version of the paper* and *to examine physical differences*. These are perhaps better regarded as formulaic expressions rather than conventional collocations, with the last three arguably including two collocations each rather than one.

The analysis of the Not Sure responses showed that the students varied widely in their use of this option. Those with a larger vocabulary size were apparently able to make more confident judgements about collocational expressions, as shown by their smaller mean number of Not Sure responses and the strong negative correlations with their scores on Part B and the Depth Test as a whole. This finding probably reflects the fact that, as learners know more words, they become more familiar with the lexical items that these words are likely to co-occur with. In fact, Nguyen and Webb's (2017) study with Vietnamese EFL learners found that as these learners' knowledge of the most frequent 3,000 words increased, their knowledge of collocations made up of these words increased accordingly, and that node word frequency was the strongest factor predicting the receptive knowledge of collocations.

In Part C of the Depth Test (Word parts), the factors influencing the learners' responses were similar to those in Part B. There was reduced use of the Not Sure option overall, reflecting the test-takers' greater confidence in identifying inflected and derived forms of nouns that they were familiar with, as indicated by their correct responses to the stem forms in Part A and the Translation Test. On the other hand, they were less confident about rejecting incorrect or inappropriate derived forms of words, even when they showed evidence of knowing the meaning of the stem form.

We should point out that there was considerable individual variation in the use of Not Sure, even in the most proficient 2K+ group. In Part B 29 (62%) of these learners chose the option 5 times or fewer, whereas four of them selected it 13-19 times. In Part C there was more use of Not Sure by 2K+ learners overall, but the tallies again showed a wide range from 10 learners who did not make use of the option at all to three who chose it 21 or 22 times.

While the Depth Test is an experimental measure, its results still provide useful insights into the participants' depth of academic vocabulary knowledge. Among the three aspects of vocabulary knowledge measured by the Depth Test, knowledge of synonyms was the best known, followed by inflected/derived forms. Collocation was the least known. Analysis of each sub-group of participants revealed that this overall pattern holds true for those who had mastered at least the most frequent 1000 words (1K and 2K+ learners). For the learners who were yet to master the most frequent 1000 words (<1K learners), although their knowledge of collocations was the smallest, their knowledge of synonyms and word parts was fairly similar. This finding may reflect these <1K learners' generally limited vocabulary knowledge. To some extent, this study is in line with previous studies which showed that knowledge of collocations and derived forms lagged behind knowledge of the form-meaning link (e.g., Chui, 2006; González-Fernández and Schmitt, 2019). However, while González-Fernández and Schmitt found that word part knowledge was harder to acquire, the present study showed that collocations were the hardest to master. There are two possible reasons for the different findings. First, González-Fernández and Schmitt measured only knowledge of derived forms, whereas this study also measured knowledge of inflected forms. Second, González-Fernández and Schmitt examined knowledge of general vocabulary while the present study investigated academic vocabulary knowledge.

VI. Conclusion

The Depth Test produced some promising results, but the formats need further investigation, in keeping with Schmitt et al.'s (2020) recommendation that tests should not be released for wider use until they have been extensively validated. Our primary purpose was to explore new ways of developing tests of depth of vocabulary knowledge that are sensitive to the needs of particular populations of learners. This was evident in the way that we defined the relevant domain of vocabulary as being the 1st 200 lemmas of the AVL, coupled with the 1st 2000 word families of the BNC/COCA lists, in order to diagnose depth of knowledge of largely known vocabulary.

As the test evolved, it consisted entirely of selected-response rather than constructed-response test formats, in response to feedback from students in an initial trial. One significant issue with selected-response items is their susceptibility to the effects of guessing behaviour by the test-takers (Stoeckel et al., 2021). We addressed the issue to some degree by adopting another request from the trials: to include the Not Sure option in Parts B and C. Our analyses

showed that the students made liberal use of this option in ways that were consistent with their fairly conservative approach to claiming knowledge of academic vocabulary items. It gives us some confidence that this population of learners tended to under-report rather than over-report their vocabulary knowledge. However, further research is needed to better understand the role of the Not Sure option in the test results and to explore individual differences in use of the option.

Another feature of the test items in Parts B and C was the inclusion of incorrect or inappropriate expressions and word forms, which tended to elicit more Not Sure responses from the test-takers, as well as an even larger number of Yes responses in some cases, rather than complete rejection. This can be interpreted as validly reflecting the limitations of the students' collocational and morphological knowledge respectively. However, there is an argument that it is pedagogically inappropriate to present test-takers with anomalous forms of this kind, and from an item-writing perspective it is challenging to create items that are clearly inappropriate and yet plausible to the learners. Thus, it remains to be seen whether our approach is an effective means of assessing receptive knowledge of collocations and word parts.

The Depth Test piloted in this study offers useful evidence on the depth of academic vocabulary knowledge of students in this specific context. It has the potential to help inform decisions on the placement of students in streams for their English course and to provide diagnostic information for class teachers as they plan their vocabulary teaching. As noted at the outset, if there is a case to persevere with assessing depth of vocabulary knowledge as something distinct from vocabulary breadth, it is important to try out new approaches that are tailored to the needs of particular populations of learners and that are adequately validated.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Akbarian, I. (2010). The relationship between size and depth for ESP/EAP learners. *System, 38(4)*, 391-401.

Alderson, J.C. (2005). *Diagnosing foreign language proficiency*. Continuum.

Anderson, R.C., & Freebody, P. (1981). Vocabulary knowledge. In J.T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-177). International Reading Association.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing 27(1)*, 101-118.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Chapelle, C.A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). Routledge.

Chui, A.S.Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching, 16*, 1-23.

Coxhead, A. (2018). *Vocabulary and English for Specific Purposes research: Quantitative and qualitative perspectives*. Routledge.

Dang, T. N. Y. (2018). The nature of vocabulary in academic speech of hard and soft sciences. *English for Specific Purposes, 51* (1), 69–83.

Dang, T. N. Y. (2020). High-frequency words in academic spoken English: Corpora and learners. *ELT Journal, 74*(2), 146–155. https://doi.org/10.1093/elt/ccz057

Dang, T. N. Y., & Webb, S. (2020). Vocabulary instruction and the good language teachers. In C. Griffiths, Z. Tajeddin, & A. Brown (Eds.), *Lessons from good language teachers* (pp. 203–218). Cambridge University Press.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing 30(2)*, 253-272.

Eyckmans, J. (2009). Towards an assessment of learners' receptive and productive syntagmatic knowledge. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 153-170). Palgrave Macmillan.

Gardner, D., & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics, 35(3)*, 305-327.

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*. Published online 2 January. https://doi-org./10.1093/applin/amy057

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 139-152). Palgrave Macmillan.

Hyland, K., & Shaw, P. (2016). Introduction. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes* (pp. 1–13). London: Routledge.

Hyland, K., &Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253.

Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal, 40(1)*, 5-22.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(4), 647–672.

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language, 33 (1)*, 126-140.

McLean, S., Stewart, J., & Batty, A.O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing, 37(3)*, 389-411.

McLean, S., & Stoeckel, T. (2021). Lexical mastery thresholds and lexical units: A reply to Laufer. *Reading in a Foreign Language, 33 (2)*, 247-259.

Meara, P. (1997). Towards a new approach in modelling vocabulary learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*, pp. 109-121. Cambridge University Press.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. John Benjamins.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use* (pp. 57-78). *EUROSLA Monograph Series 2*. European Second Language Association.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines 5*, 12-25.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63(1)*, 59-82.

Nation, P. (2012). The Vocabulary Size Test. Unpublished ms. Retrieved from: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I.S.P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I.S.P. (2020). The BNC/COCA Level 6 word family lists. Unpublished paper. Available at https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf

Nation, I.S.P., & Coxhead, A. (2021). *Measuring native-speaker vocabulary size*. John Benjamins.

Nguyen, L.T.C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal 42(1)*, 86-99.

Nguyen, T.M.H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research, 21(3)*, 298-320.

Pearson, P.D., Hiebert, E.H., & Kamil, M. (2017). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly, 42(2)*, 282-296.

Pecorari, D., Shaw, P. & Malmström, H. (2019). Developing a new academic vocabulary test. *Journal of English for Academic Purposes, 39(1),* 59-71.

Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56(2)*, 282-308.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19* (2), 12-25.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10* (3), 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Erlbaum.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209-227). Benjamins.

Read, J. (2013). Reflecting on the validity of vocabulary assessments. Paper presented at the Vocab@Vic Conference, December 2013, Victoria University of Wellington.

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research 21(1)*, 12-30.

Schmitt, N. (1998). Tracking the incidental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48(2),* 281-317.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning 64(4)*, 913-951.

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53(1)*, 109-120. doi:10.1017/S0261444819000326

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47(4),* 484 - 503.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18(1)*, 55–88. doi: 10.1177/026553220101800103

Schmitt, N., & Zimmerman, C.B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36(2)*, 145-171.

Stoeckel, T., Bennett, P., & McLean, S. (2016). Is "I Don't Know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly,50,* 965-975. doi: 10.1002/tesq.325

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43 (1)*, 181-203. doi:10.1017/S027226312000025X

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System, 93*, 1-13.

Verhallen, M., & Schoonen, R. (1993). Word definitions of monolingual and bilingual children. *Applied Linguistics, 14(4)*, 344–365.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27(1),*33-52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28(1),* 46-65.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168(1)*, 34-70.

Wesche, M.B., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: depth vs. breadth. *Canadian Modern Language Review 53(1)*, 13-39.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System, 35(3)*, 182-191.

Zhang, D., & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: Issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*(1). Retrieved from: https://ore.exeter.ac.uk/repository/handle/10871/28216.

Zhang, D., & Yang, X. (2016). Chinese L2 learners' depth of vocabulary knowledge and its role in reading comprehension. *Foreign Language Annals, 49(4)*, 699-715.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension. *Language Teaching Research.* Advanced online publication. https://doi.org/10.1177/1362168820913998

Zhao, P., & Ji, X. (2016). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal, 49(3)*, 308-321.

NOTES

[1] The university entrance English exam and the VSTEP are standardized tests. These tests are designed following Vietnam's six-level framework of foreign language proficiency, which was adapted from the CEFR. In other words, Levels 1-6 in this framework correspond to the A1-C2 levels on the CEFR.

[2] It should be noted that "synonym" is used in a loose sense here to refer to a word that is semantically related to the target word or represents one aspect of its meaning.

[3] In Ishii and Schmitt's (2009) study of Japanese learners, which employed a constructed-response format to measure knowledge of derived forms, the mean score was much lower than for the other three tests: 37% vs. 64-66%.

[4] In relation to vocabulary tests in particular, McLean (2021) argues that cut scores should be set on the basis of empirical evidence of their appropriateness for the assessment purpose, such as selecting materials for different uses in the classroom.

[5] A reviewer suggested that this might be explained by the fact that the 3000-word level scores were the most reliable. However, apart the 1000-word level, with a coefficient of .83, the reliabilities of the other four levels were very comparable, at .925, .929, .925 and .935 respectively.

[6] Pecorari, Shaw & Malmström (2019) have subsequently published such a test.